DOCUMENT RESUME

ED 107 629 95 SP 009 252

AUTHOR Winne, Philip H.
TITLE A Critical Review of Experimental Studies of Teacher Questions and Student Achievement.
INSTITUTION Stanford Univ., Calif. Stanford Center for Research and Development in Teaching.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE [75]
CONTRACT NE-C-00-3-0061
NOTE 21p.

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Academic Achievement; Affective Behavior; *Educational Experiments; Educational Research; Literature Reviews; Questioning Techniques; Research Design; Research Methodology; *Research Problems; *Teacher Influence; *Teaching

ABSTRACT

The purpose of this paper is to summarize and evaluate experiments which examined the effects of teacher questions on student achievement. The studies reviewed are of two types: (a) training experiments, in which the independent variable is teacher training; and (b) skills experiments, in which the frequency and manner of use of a teaching skill is prescribed by the experimenter. The first section of this paper presents brief overviews of both training and skills experiments. Each overview lists (a) grade, (b) subject, (c) independent variable, (d) dependent measure, (e) teaching time, (f) analysis and results, (g) comments, and (h) conclusions. The second section discusses the experiments and presents suggestions for improving the quality of research on teaching. These suggestions include the following areas: (a) reporting the study, (b) design, (c) analysis, (d) dependent measures, and (e) general questions of method. The last section presents conclusions gathered from the studies reviewed and warns of misleading research supported by superficial claims of valid methodology. (PB)

ED107629

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A CRITICAL REVIEW OF EXPERIMENTAL STUDIES OF TEACHER QUESTIONS AND STUDENT ACHIEVEMENT[1]

Philip H. Winne
Stanford Center for Research and Development in Teaching
Stanford University

There can be little doubt that teacher questions are assumed to be important factors influencing student achievement. Indeed, in a previous review of the topic, Gall (1970) labeled this statement a "truism." In the last decade, efforts have been made to test this belief with experiments. The purpose of this paper is to summarize and evaluate the experiments which examined the effects of teacher questions on student achievement. As Heath and Nielson (1974) showed, research on teaching often exhibits flaws in method. This paper identifies such errors in the domain of research on teacher questions and provides some suggestions for improving research on teaching in general.

The studies reviewed here are of two kinds. Training experiments are studies in which the independent variable is merely teacher training. Following training, teachers are free to use the skill(s) on which they were trained at their discretion in teaching. Since the skill(s) are not necessarily used with the same frequency or in the same manner by different teachers within a treatment group, it is incorrect to label the skill itself the independent variable used in these studies.

In contrast, skills experiments are studies in which the frequency and manner of use of a teaching skill is prescribed by the experimenter. Thus, the teaching skill is the nominal independent variable in these studies. It may not be the actual independent variable, however, if teachers within a treatment group who should be following the experimentally prescribed use of the teaching skill vary in their actual delivery of the teaching act.

The next two sections present very brief overviews of training experiments and skills experiments, respectively. The last two sections present suggestions for improving the quality of research on teaching and a summary of my conclusions gathered from the studies reviewed here about the effects of teacher questions on student achievement.

---

[1]The research reported herein was conducted at the Stanford Center for Research and Development in Teaching, which is supported in part by the National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed in this draft do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. NE-C-00-3-0061.)

SP009 252

Training Experiments

Beseda (1972)

Grade: 9-12 (mixed grade classrooms)

Subject: American history, world history, U.S. government

Independent Variable: Intern training plus feedback on classroom performance for convergent and divergent questions vs. no training or feedback.

Dependent Measure: Iowa Test of Educational Development, Ability to Interpret Reading Materials in the Social Studies (80 items, multiple choice), Sequential Tests of Educational Progress, Social Studies (70 items, multiple choice); Watson-Glaser Critical Thinking Appraisal (100 objective items).

Teaching Time: ___ 50 minute lessons over 8 weeks.

Analysis & Results:

1. ANCOVA of ITED showed no treatment differences ($F_{1,425} = 3.09$, $p > .05$).
2. ANCOVA of STEP showed no treatment differences ($F_{1,403} = 1.81$, $p > .05$).
3. ANCOVA of Watson-Glaser showed no training > trained group ($F_{1,401} = 5.17$, $p < .05$).
   NOTE: parallel pretest served as the covariate in each case.

Comments:

1. Analysis of the observational data from a single lesson at the end of 8 weeks of training is not representative of teaching over the full treatment period.
2. ANCOVA assumptions are not mentioned, especially homogeneity of regression.
3. Students as units of analysis are not independent sampling units.
4. Standardized measures of achievement are designed for high stability coefficients and, thus, are not likely to reveal treatment differences.
5. Reanalysis of observational data using the mean number of types of questions (vs. author's use of total number of types of questions) shows no relation between treatment and teacher acts in the single lesson observed.
6. There probably are some differences on the ITED measure, $p = .079$ (reported by author) assuming his analysis acceptable.

Conclusions:

1. The treatment delivered to students is unknown.
2. Analyses are inaccurate.
3. There is a poor choice for the dependent measures.
4. The reported results are not valid.

Lynch et al.: Study A

Grade: 1-3 (mixed grade lesson groups)

Subject: Science, "why birds sing"

Independent Variable: Interns were told to teach for factual recall vs. concept mastery in a 1 hour orientation meeting.

Dependent Measure: 12 item recall test (KR_{20} = .38); 10 item concept mastery test (KR_{20} = .52).

Teaching Time: One 15-30 minute lesson.

Analysis & Results: MANOVA showed significant differences for teacher objectives ($F_{4,58} = 4.20$, $p < .025$) with recall group > concept mastery group on recall test, and significant differences for grade ($F_{2,30} = 9.12$, $p < .001$) with grades 3 and 2 > grade 1 on concept mastery test.

Comments:

1. Analysis of the observational data showed that the concept mastery group of interns asked significantly fewer knowledge questions, significantly more higher cognitive questions, but no interrater agreement coefficient was given.
2. Cell sizes for MANOVA are in the ratio 1 : 1 : 1.25 : 1.50 : 2 : 2.25 with largest call variances in the ratio 1 : 7.3 (recall test), 1:14 (concept mastery test).
3. Dependent measures have very low reliabilities.
4. Teaching delivered to students is ill-defined, very short.

Conclusions:

1. The treatment delivered to students is unknown.
2. Analyses are inaccurate.
3. The dependent measure is not trustworthy.
4. Teaching time is insufficient for meaningful results.
5. Reported results are not valid.

Lynch et al.: Study B

Grade: 4-6 (mixed grade lesson groups)

Subject: Symbol recognition for artificial code.

Independent Variable: Interns were told to teach for factual recall vs. concept mastery in a 1 hour orientation meeting.

Dependent Measure: 24 item knowledge-recall test (KR_{20} = .79); 33 item concept mastery test (KR_{20} = .56).

Teaching Time: One 30-40 minute lesson.

Analysis & Results: MANOVA showed significant differences for teacher objectives ($F_{1,30} = 9.46$, $p < .001$) with recall group > concept mastery group on knowledge-recall test ($F_{1,30} = 17.52$, $p < .001$), and significant differences for grade ($F_{2,30} = 2.76$, $p < .05$) with grade 6 > 5 > 4 on knowledge-recall test ($F_{2,30} = 6.19$, $p < .01$).

Comments & Conclusions:

1. Since study B was a methodological twin of study A, the same comments and conclusions apply.

Millett (1967)

Grade: 8-12

Subject: Social studies, "the McCarthy hearing"

Independent Variable: Interns were trained by oral discussion vs. video model vs. oral discussion plus video model vs. no trainirg.

Dependent Measure: 12 item short essay test (split half coefficient = .82).

Teaching Time: One 9-23 minute lesson.

Analysis & Results: ANOVA of posttest showed no differences ($F_{3,30} = 0.83$, $p > .05$).

Comments:

1. Analyses of observational data are slightly misleading since some teachers who did not give the achievement test were included in observational data.
2. Testing time ranged from 7 minutes to 20 minutes, with a mean of about 12 minutes. Thus, the test is relatively speeded and gives unequal opportunity for students to show what they learned.
3. Variation in teaching time makes comparisons across groups difficult since treatments probably varied as a function of time for the lesson.

Conclusions:

1. Treatment differences between groups are relatively unknowr.
2. Variable teaching time plus variable testing time make comparisons of lesson groups within treatments and between-group treatment comparisons untrustworthy.
3. Reported results are not valid.

Rogers & Davis (1971)

Grade: 5

Subject: Social studies, The West Indies

Independent Variable: Intern training on asking higher cognitive questions vs. no training.

Dependent Measure: 35 item multiple choice test (unspecified reliability coefficient = .75); 5 items for each of seven categories of questions from Sanders also considered as separate subscales.

Teaching Time: Four 35-40 minute lessons.

Analysis & Results: ANOVA showed no significant differences for total test ($F_{1,531} = 2.71$), memory subtest ($F_{1,531} = 2.01$), translation subtest ($F_{1,531} = .00$), interpretation subtest ($F_{1,531} = .10$), application subtest ($F_{1,531} = .32$), synthesis subtest ($F_{1,531} = 1.39$), and evaluation subtest ($F_{1,531} = .00$); significant treatment differences for analysis subtest ($F_{1,531} = 14.77$) with untrained group > trained group.

Comments:

1. Seven separate analyses of observational data performed for each category of questions are not independent since (a) the same sample of teachers is used for each analysis and (b) the data were proportions so that the sum of seven proportions must total 100%.
2. Measuring teachers' use of questions by proportions may be misleading; the largest absolute difference for a type of question may be less than 2 questions per lesson if teachers asked 20 questions per lesson.
3. The analyses incorrectly use students as the unit of analysis since they are not independent units in this design.
4. The reliability of 5 item subtests is very low (roughly .12 by the Spearman-Brown formula, but it is likely this figure is slightly misleading).

Conclusions:

1. Analyses are inaccurate.
2. Treatment variation within experimental groups makes comparison across groups difficult.
3. Subtest analyses are not trustworthy.
4. Reported results are not valid.

Skills Experiment

Aagaard (1973)

Grade: 11

Subject: Chemistry, radioactivity and radiation

Independent Variable: Scripted lessons with 250 higher cognitive questions vs. 310 knowledge questions vs. no teacher initiated questions.

Dependent Measure: 45 item multiple choice test (KR_{20} = .88)

Teaching Time: Ten 60 minute lessons over two weeks.

Analysis & Results:

1. ANOVA of pretest (same measure as posttest) showed no differences ($F_{2,734} = 2.31$, $p > .05$).
2. ANOVA of posttest showed a significant treatment effect ($F_{2,734} = 8.30$, $p < .01$); Scheffé contrasts showed knowledge < higher cognitive questions.
3. ANOVA of gain score showed a significant effect ($F_{2,734} = 6.98$, $p < .01$); Scheffé contrasts showed knowledge questions < higher cognitive questions ($p < .05$), mean of no questions plus knowledge questions < higher cognitive questions.
4. ANOVA of unspecified IQ measure showed significant differences ($F_{2,734} = 3.74$, $p < .05$).
5. Multiple regression analysis of gain score showed a significant increase in R^2 for the model with groups plus IQ vs. groups only (increase $R^2 = .031$, $F_{1,732} = 23.99$, $p < .01$).
6. A priori contrasts of gain score residualized on IQ showed no questions < higher cognitive questions ($F_{2,733} = 4.80$, $p < .01$), knowledge questions < higher cognitive questions ($F_{2,733} = 13.40$, $p < .01$), and mean of no questions plus knowledge questions < higher cognitive questions ($F_{2,733} = 11.10$, $p < .01$).

Comments:

1. Scheffé contrasts from the ANOVA of pretest showed no questions > knowledge questons ($p < .01$), no questions > higher cognitive questions ($p < .01$).
2. Scheffé contrasts from the ANOVA of IQ showed no questions > knowledge questions ($p < .03$).
3. Differences favoring the no questions group for both pretest and IQ measures suggest that a finding of no differences on posttest would show that the treatment had an effect.

4. Observational data is inadequate; only 8 of 30 lessons or 27% of teaching time was observed. These observations were sampled unsystematically.
5. Students is not the correct choice for the unit of analysis since neither students nor classrooms were randomly assigned to treatments.
6. Analyses using gain scores are unreliable.
7. No correlation between IQ and posttest is given.
8. A mean gain of only 7 items over 10 lessons suggests that teaching, the curriculum, or some other factor inhibited the effectiveness of the lessons.

Conclusions:

1. Analyses are inaccurate.
2. The treatment delivered to students is relatively unknown.
3. Reported results are not valid.

Buggey (1971)

Grade: 2

Subject: Social studies; one unit on rules and a second unit on locations.

Independent Variable: 70% higher cognitive questions vs. 30% higher cognitive questions vs. no instruction on curriculum; sex; urban vs. suburban school location.

Dependent Measure: Sum of scores from two 30 item multiple choice tests on each of the two units (KR_{20} = .84 for summed scores).

Teaching Time: 8 ___ minute lessons over 3 weeks on unit 1, and 8 ___ minute lessons over 3 weeks on unit 2.

Analysis & Results: ANOVA of posttest showed significant differences for teaching method ($F_{2,96} = 269.99$, $p < .01$) and school location ($F_{1,96} = 10.89$, $p < .01$); Neuman-Keuls contrasts showed 70% higher cognitive questions > 30% higher cognitive questions and no instruction ($p < .01$ for both), 30% higher cognitive questions > no instruction ($p < .01$).

Comments:

1. Since KR_{20} is a measure of internal consistency, it is curious that a 60 item test composed of two supposedly different subscales measuring different content has such a large coefficient.
2. There was no observation of teaching.
3. Since the first unit was taught by one teacher and the second unit by a different teacher, results may reflect a topic by teacher interaction or warm-up effect.
4. Using a no instruction group as control wastes resources. The threat of reactive testing effects by administering a pretest seems minor.

Conclusions:

1. Analyses may confound treatment differences with a teacher by topic interaction.
2. The treatment delivered to students was not observed, but since the experimenter was one teacher (the other two were also Ph.D. candidates doing a replication or extension of this design), it seems relatively safe to assume the treatment was known. It was neither analyzed nor reported, however.
3. Results probably are valid as reported.

Church (1970)

Grade: Standard 4 (approximately Grade 10-11)

Subject: Science, electricity

Independent Variable:

1. Study A: 171 primary questions vs. 53 primary questions.
2. Study B: 65% open primary questions for 110 minute (long) lessons vs. 65% open primary questions for 66 minute (short) lessons vs. 35% open primary questions for 70 minute (short) lessons.
3. Study C: Teacher response to secondary questions: prompts vs. extensions vs. teacher gives answer.
4. Study D_1: Number of questions, Q and Q/2 (actual number not specified).
5. Study D_2: Number of questions, 171 primary questions vs. "reduced as far as possible" (actual number not specified).

Dependent Measure: Achievement test corrected score (correction measure and method unspecified).

Teaching Time:

1. Studies A, D_1, D_2: 3 lessons of varying length in minutes.
2. Studies B, C: 4 lessons of varying length in minutes.

Analysis & Results:

1. Study A: 171 primary questions ($\bar{X} = 36.3$) > 53 primary questions ($\bar{X} = 31.9$).
2. Study B: 65% open questions long lesson ($\bar{X} = 31.4$) and 35% open questions short lesson ($\bar{X} = 31.2$) > 65% open questions short lesson ($\bar{X} = 27.9$).
3. Study C: prompting ($\bar{X} = 31.4$) > extension ($\bar{X} = 29.0$) ≐ teacher gives answer ($\bar{X} = 27.9$).
4. Study D_1: Q questions ($\bar{X} = 39.8$) = Q/2 questions ($\bar{X} = 38.0$).
5. Study D_2: 171 primary questions ($\bar{X} = 36.3$) > "reduced as far as possible" ($\bar{X} = 33.4$).

Comments:

1. The measure and method used for correcting achievement test scores are not specified.
2. There is no presentation of basic statistical information, e.g., unit of analysis, standard deviations, sample size, inferential tests of hypotheses.
3. The study used a "middle group" of students from classrooms as its sample; this restricts the range of individual differences and limits generalizability due to unrepresentativeness of the sample.
4. For every corrected mean score in all the studies, the greater (greatest) mean is associated with lessons taking longer time; the differences in average time for lessons range from 1 minute to 39 minutes.

Conclusions:

1. The reported results contribute little to knowledge about the effects of teacher questions on student achievement.

Martikean (1973)

Grade: 3-4

Subject: Science, plants and seeds

Independent Variable: 107 higher cognitive questions plus 9 knowledge questions vs. 5 higher cognitive questions plus 52 knowledge questions.

Dependent Measure: 11 item objective achievement test.

Teaching Time: Not specified, presumably 1 lesson.

Analysis & Results:

1. t-test on parallel pretest means showed no difference ($t_{29} = .07$, $p > .05$).
2. t-test on posttest means showed no difference ($t_{29} = .21$, $p > .05$).
3. t-test on mean gain scores showed no difference ($t_{29} = .14$, $p > .05$).

Comments:

1. The absence of basic statistical information, e.g., standard deviations, reliability coefficients, limits interpretation.
2. No observation of teaching.
3. The t-test on gain scores is unreliable; the analysis should have been a t-test for correlated samples.
4. The ratio of questions is approximately 2:1. This suggests that time varied considerably.

Conclusions:

1. The treatment delivered to students is unknown.
2. The results are relatively uninterpretable due to poor reporting.

Ryan (1973)

Grade: 5

Subject: Social studies, geography

Independent Variable: 75% higher cognitive questions vs. 5% higher cognitive questions vs. no instruction on curriculum.

Dependent Measure: 58 item knowledge-recall multiple choice test (KR_{20} = .89), 46 item higher cognitive question multiple choice test (KR_{20} = .86).

Teaching Time: 9___ minute lessons over 2 weeks.

Analyses & Results:

1. ANOVA of knowledge questions posttest showed significant treatment effect ($F_{2,103}$ = 21.37, $p < .01$); Neuman-Keuls contrasts showed 75% higher cognitive questions and 5% higher cognitive questions > no instruction ($P < .01$ for both).
2. ANOVA of higher cognitive questions posttest showed significant treatment effect ($F_{2,103}$ = 5.70, $p < .01$); Neuman-Keuls contrasts showed 75% higher cognitive questions > no instruction ($p < .01$).
3. ANOVA of knowledge questions retention test showed significant treatment effect ($F_{2,103}$ = 16.15, $p < .01$); Neuman-Keuls contrasts showed 75% higher cognitive questions and 5% higher cognitive questions > no instruction ($p < .01$ for both).
4. ANOVA of higher cognitive questions retention test showed significant treatment effect ($F_{2,103}$ = 5.64, $p < .01$); Neuman-Keuls contrasts showed 75% higher cognitive questions > no instruction ($p < .01$).

Comments:

1. The author states teachers were observed occasionally, but presents no data on their adherance to treatment.
2. Each treatment was delivered by only one teacher; treatment effects are confounded with teachers and can be attributed to treatments, teachers, or a treatment by teacher interaction.

3. The author misinterprets the data; unreliable differences ($p > .10$) are claimed to show consistent effects.
4. There is insufficient information about how long students had to respond to a total of 104 multiple choice items. This raises the question of test speededness. If the tests were speeded, KR_{20} coefficients are spuriously high and the analyses lack power.
5. A no instruction group is a waste of resources.

Conclusions:

1. There is confounding of treatment with teacher.
2. The dependent measure may be unreliable.
3. The results are not interpretable regarding the effects of questions. The only differences are between students who studied the curriculum they were tested on and those who didn't study this curriculum.

Ryan (1974)

Grade: 5

Subject: Social studies, geography

Independent Variable: 75% higher cognitive questions vs. 5% higher cognitive questions vs. no instruction on curriculum.

Dependent Measure: 58 item knowledge-recall multiple choice test (KR_{20} = 89), 46 item higher cognitive question multiple choice test (KR_{20} = .86).

Teaching Time: 9 ___ minute lessons over two weeks.

Analyses & Results:

1. ANOVA of knowledge questons posttest showed significant treatment effect ($F_{2,104} = 36.03$, $p < .01$); Neuman-Keuls contrasts showed 75% higher cognitive questions and 5% higher cognitive questions > no instruction ($p < .01$ for both).
2. ANOVA of higher cognitive questions posttest showed significant treatment effect ($F_{2,104} = 5.24$, $p < .01$); Neuman-Keuls contrasts showed 75% higher congnitve questions > no instruction ($p < .01$), and 5% higher cognitive questions > no instruction ($p < .05$).

3. ANOVA of knowledge questions retention test showed significant treatment effect ($F_{2,104} = 20.87$, $p < .01$); Neuman-Keuls contrasts showed 75% higher cognitive questions and 5% higher cognitive questions > no instruction ($p < .01$ for both).
4. ANOVA of higher cognitive questions retention test showed significant treatment effects ($F_{2,104} = 7.45$, $p < .01$); Neuman-Keuls contrasts showed 75% higher cognitive questions and 5% higher cognitive questions > no instruction ($p < .01$ for both).

Comments:

1. Since this study is a methodological twin of Ryan's 1973 study, the same comments apply.
2. The df in the report are consistently 1 less than they should be for MS_e.
3. The author states incorrectly that all three groups were run concurrently; the control group for 1974 was the same as that for 1973 (confirmed by personal communication, F. Ryan, February 20, 1975).

Conclusions:

1. The results do not contribute to knowledge about the effects of questions.

Savage (1972)

Grade: 5

Subject: Social studies, one unit on rules and a second unit on locations.

Independent Variable: 70% higher cognitive questions vs. 30% higher cognitive questions vs. no instruction on curriculum; sex; urban vs. suburban school location.

Dependent Measure: Sum of scores from two 30 item multiple choice tests ($KR_{20} = .84$ for summed scores).

Teaching Time: 8 ___ minute lessons over 3 weeks on unit 1; 8 ___ minute lessons over 3 weeks on unit 2.

Analysis & Results: ANOVA of posttest showed significant differences for teaching method ($F_{2,84} = 80.84$, $p < .01$) and sex ($F_{1,84} = 14.95$, $p < .01$) with females > males. Neuman-Keuls contrasts showed 70% higher cognitive questions and 30% higher cognitive questions > no instruction ($p < .01$ for both).

Comments & Conclusions:

1. Since this study is a methodological twin of Buggey's (1971), the same comments and conclusions apply. The reported results are probably valid.

Tyler (1971)

Grade: 2

Subject: Social studies, one unit on rules and a second unit on locations.

Independent Variable: Teacher asked questions vs. students read questions (70% higher cognitive and 30% knowledge questions were identical for both) vs. no instruction on curriculum; sex; urban vs. suburban school location.

Dependent Measure: Sum of scores from two 30 item multiple choice tests (KR_{20} = .84 for summed scores).

Teaching Time: 8 ___ minute lessons over 3 weeks for unit 1; 8 ___ minute lessons over 3 weeks for unit 2.

Analysis & Results: ANOVA of posttest showed significant differences for teaching method ($F_{2,108} = 121.95$, $p < .01$); school location ($F_{1,108} = 97.40$, $p < .01$), with suburban > urban; treatment x school location ($F_{2,108} = 4.97$, $p < .05$); and sex x school location ($F_{2,108} = 7.23$, $p < .01$); Neuman-Keuls contrasts for teaching method showed teacher asked questions and student read questions > no instruction ($p < .01$ for both), teacher asked questions > student read questions ($p < .05$).

Comments & Conclusions:

1. Since this study is a methodological twin of Buggey's (1971), the same comments and conclusions apply. The reported results are probably valid.

Discussion

The four training studies and eight skills studies examined in this review highlight several methodological flaws that are probably common to much research on teaching. The correction of these flaws and more astute consideration of the limitations of method should become prominent in future investigations. These issues are briefly summarized in the following.

Reporting the Study. Several of the studies suffer from inadequate reporting. Training of teachers should be described briefly so that meaningful comparisons can be made of studies which examine the same or similar variables but use different training methods to get teachers to use the teaching actions under investigation. Reference should be made to documents used in training, the length of training, and standardized training exercises such as microteaching. Of particular importance is the need to explicitly and exhaustively describe the teaching behaviors trained as well as those untrained. I recommend that separate reports fully describing the training be cited and made available. This will have the triple benefit of saving space in journal articles, of fully describing the independent variable, and of contributing to knowledge about the effectiveness of various training techniques for particular teaching acts.

A second limitation of reporting obvious in several studies is the insufficient presentation of descriptive statistics, including standard deviations, reliability coefficients, and correlations between measures used as covariates or residualizing variables and posttest scores. These statistics contribute much to permitting a reader to form his own interpretation of the results.

Design. This review of experimental studies in a limited area of teacher effectiveness corroborates the finding of Health and Nielson (1974) that many research efforts are methodologically inadequate. All research on teacher effectiveness should include observation of the teaching that takes place. Without this component, the actual treatment of an experiment is unknown to a degree that casts serious doubt on the validity of inferences drawn from the data. Furthermore, a simple statement that teachers adhered to the definition of the treatment is insufficient. The degree of variation from the treatment as well as the characteristics of the variation(s) may be critical in judging what produced the observed results. Therefore, research studies should include a description of variations in the treatment and, where space permits, a formal analysis of the degree of variation should be presented. In the absence of available space, a citation to a document containing such analyses should be available to supplement a summary presentation of the analyses in the published paper.

Studies should be designed to allow an estimate of the variance in the dependent measure attributable to variation between the teachers. This seems accomplished most easily by making teachers a factor in the experimental design that is fully crossed with treatments. Failure to include this factor will usually leave treatment effects fully or partially confounded with teacher effects, thus confusing the interpretation of why the results turned out as they did.

The arguments over whether to use classrooms (or teachers) or individual students as the unit of analysis are complex. The simplest resolution of the choice is the following. In analyses like multiple regression, analysis of variance, analysis of covariance, and the like, there must not be a systematic relation between the units of analysis and any factor in the design to avoid the problems of correlated errors. Assignment of an intact classroom to a treatment and then using students as the units of analysis probably violates this dictum for almost any experimental factor. At the least, there probably exist patterns of social interation within an intact classroom that influence student response tendencies and attentional factors. Since the ideal condition of randomly assigning all students to a treatment is seldom feasible, researchers must find a middle ground best suited to the questions posed in their studies. A powerful control for classroom is to randomly divide each classroom into equal halves, thirds, or quarters. These randomly formed groups then should be randomly assigned to treatments that differ only on one dimension of the experimental design. For example, a study examining the relative effects of knowledge and higher order questions should randomly halve classrooms and randomly assign each half to a level, knowledge questions or high order questions, of the experimental factor of type of questions. Any other independent variables should be held constant for the two groups formed from the same classroom. This assignment procedure can validly use students as the unit of analysis under the condition that teachers who taught each half were statistically and reasonably equivalent in their rendition of the treatment. This is because the half-classrooms differ randomly with respect to the independent variable which has been varied. Since all other independent variables are constant over the two halves, and since there has been random assignment to the experimental factor varied, the likelihood of correlated errors is reduced considerably. The generalization of this reasoning to an independent variable with more than two levels is straightforward.

In addition to the caveat that teachers in each half classroom be equivalent, two other cautions must be heeded. Aggregating information over classrooms requires that classrooms be randomly assigned to treatment conditions under the restriction outlined above. It also requires that where more than one classroom is assigned to the same treatment condition, great care be taken that the classrooms are not systematically different insofar as possible. This demands rigorous investigation of the sample characteristics at the level of classrooms. For example, classrooms should be compared on a measure of general ability like vocabulary, an interest or attitude survey relating to the teaching variables to be studied and the experimental curriculum, or other measures relevant to the particular investigation.

The second caution in this method of attacking the problem of units of analysis is one pertaining to sample size. Dividing a classroom of 30 students into a large number of groups yields only a small number of students per group. This can destroy the statistical power gained by the sampling procedure outlined above. There seem to be only two alternatives to the dilemma of wanting to ask many research questions with limited resources: ask fewer questions or sacrifice the probability of identifying true differences due to decreases in power resulting from small sample sizes.

The latter option can be disastrous for building knowledge about teachers' ability to influence student learning since we play a dart game with a small board and unfeathered darts to begin with. The former option is to be valued. It requires that each study ask a few piercing questions and that sets of studies be programatic so that the range of information desired can be obtained over several investigations.

Analysis. The studies reviewed here reveal both errors of omission and comission regarding the appropriate use of statistical analyses. Perhaps most obvious in almost every study is the absence of an explicit statement that critical assumptions underlying statistical analyses were tested and judged valid. Moreover, it was shown that several important assumptions were not justified in some studies. A preliminary test for the validity of assumptions is essential for good research which relies on statistical analyses for inferences of the causal effects of an independent variable. Every study ought to state that assumptions were examined and how they were examined. The presence of two to six such sentences would greatly improve the interpretability of research on teaching.

I also recommend that investigators report a measure of the proportion of variance in the dependent measure that can be accounted for by each or several of the independent variables. This measure need not be ω^2 (see Hays, 1973), but this statistic or a suitable equivalent can be very informative about how influential the treatment is in determining values of the dependent variable. A treatment that exhibits a low value of ω^2 or an equivalent statistic is not a major determinant of scores on the dependent variable. It might be said that such a treatment is not pure or that the effects observed depend on other unidentified factors of their interactions with the independent variable manipulated in the research. Thus, such treatments need dissection or further consideration in future research before they can be accepted as the causal agent promoting the differences observed between treatment groups that are statistically different at a given level of significance.

Finally, where continuous variables such as prior achievement, general intelligence, and the like are used as "controlling" variables, these variables should enter the analyses as they are, not as blocking variables. Analyses which use a median-split or tripartite blocking on continuous data lose considerable statistical power (see Cronbach & Snow, in press). It is a much better arrangement to use a general linear model or generalized regression analysis in which the continuous variables are forced to be the first variable used to partition variance in the dependent measure. These procedures are further described by Walberg (1971) and Cohen (1968).

Dependent Measures. Several points about necessary characteristics of measures of student achievement also have been shown to have great influence in judging the quality of research on teaching. The reliability of an achievement test is a key factor in the faith which can be placed in differences observed between treatments since a low reliability indicates that students' scores are more reflective of chance variation in test responses rather than variation attributable to true ability. This statement, of course, rests on the tenets of classical measurement theory (e.g., see Gulliksen, 1950).

The reliability of a dependent measure also has an influence on the power of statistical analyses. A low reliability, i.e., large error score components, will inflate the error term in an analysis, thus decreasing the possibility of claiming differences when they may really be present.

As discussed previously, all students should have an equal opportunity to respond to all items of a test measuring learning. Furthermore, unless speed is a natural element in a particular type of school learning, measures of achievement should not be overly speeded. Not only does this tend to spuriously inflate estimates of reliability, but since students do not reach some or many test items, it also tends to decrease the content validity of the sample of items chosen to reflect the domain of instruction.

General Questions of Method. Two points unaddressed in the foregoing merit consideration in any study of research on teaching. Learning some part of a curriculum is not a short lived event. The content material and processes of thought encouraged by exposure to a curriculum via particular methods of teaching are not independent of all that students have acquired by previous experience and instruction. Relative to the knowledge and abilities which students can bring to a learning situation, a short instructional period is not likely to be particularly outstanding as an agent for changing student achievement scores. That is, of course, unless the measure of achievement focuses on little more than rote recall. The duration for instruction ought to be at least some small number of lessons, say five to ten. This will allow students some time to adjust to the style of teaching from which they are to learn. It also will allow the presentation of material that requires students to comprehend information in the sense of being able to manipulate it relative to a purpose.

A second general point is that research on teaching should be more attuned to the learning characteristics of students being taught. For example, only one study on teachers' questioning strategies (Martikean, 1973) has raised the issue of whether the distinction between knowledge questions and higher cognitive questions is appropriate at all developmental levels. The studies reviewed here show the same type of question used to teach second graders as eleventh graders. Changes in memory span, ability to organize information, and the acquisition of strategies for reasoning may be quite variant for students in the age ranges of eight to sixteen years. Yet, research seems to assume blithely that these differences are unworthy of mention, no less direct question. The question of aptitude-treatment interactions is an important one for research on teaching and should not be dismissed casually (cf. Cronbach & Snow, in press).

Conclusion

This review of experiments which examined the effects of teacher questions on student achievement was done under the assumption that the label experimental study was not sufficient proof for accepting conclusions put forth by the investigators. Intense consideration of the methodology of the twelve studies in this area showed that nine of them probably could not speak validly to the degree of influence that teacher questions have on student achievement. Of the three studies (Buggey, 1971; Savage, 1972;

Tyler, 1971) that were relatively sound methodologically, only two obtained differences for students who studied the material on which they were tested. Buggey's (1971) study suggests that higher cognitive questions lead to improved achievement relative to lower cognitive questions for second graders. Tyler's (1971) dissertation implies that questions framed by teachers are more effective than questions presented in text for second graders. Savage's (1972) failure to replicate Buggey's (1971) results in the fifth grade could result from several factors, some of which may be differences in students' level of development, the inappropriateness of the same instructional materials for one of the two grade levels, different teachers teaching different halves of the two unit curriculum, and so forth. One telling statistic is that the no study group of second graders had a mean posttest score of 15.89 while that for the fifth graders was 20.00. This suggests a large difference in prior knowledge, general reasoning ability, test wiseness, or some combination of these and other factors.

The presence of only these three studies does not provide a sturdy base for generalizations about the effects of teacher questions on student achievement. Perhaps more important, however, this review has shown that consumers of educational research need to be alert when reading studies and reviews of experimental investigations. Attributing causality is not a product easily obtained by conducting an experiment. Considerable care and expertise are required in designing, analyzing, interpreting, and reporting good research. This paper has offered several suggestions and alternatives for bettering these practices, although it is by no means definitive on these concerns. It is important that the quality of educational research be high so that neither further research nor educational practice is misled by superficial claims to strong method.

References

Aagaard, S. A. Oral questioning by the teacher: Influence on student achievement in eleventh grade chemistry (Doctoral dissertation, New York University, 1973).

Beseda, C. G. Levels of questioning used by student teachers and its effect on pupil achievement and critical thinking ability. (Doctoral dissertation, North Texas State University, 1972).

Buggey, L. J. A study of the relationship of classroom questions and social studies achievement of second-grade children (Doctoral dissertation, University of Washington, 1971).

Church, J. An experimental study of differing teaching techniques in the teaching of a science topic at the standard four level. Unpublished manuscript, University of Canterbury, New Zealand, 1970.

Cohen, J. Multiple regression as a general data-analytic system. Psychological Bulletin, 1968, 70, 426-443.

Cronbach, L. J. & Snow, R. E. Aptitudes and instructional methods. New York: Irvington Publishers, in press.

Gall, M. D. The use of questions in teaching. Review of Educational Research, 1970, 40, 707-721.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hays, W. L. Statistics for the social sciences. New York: Holt, Rinehart, and Winston, 1973.

Heath, R. W. & Nielson, M. A. The research basis for performance-based teacher education. Review of Educational Research, 1974, 44, 463-484.

Lynch, W. W., Ames, C., Barger, C., Frazer, W., Hillman, S., & Wisehart, S. Effects of teachers' cognitive demand styles on pupil learning (Final Report 30.3). Bloomington, IN: Center for Innovation in Teaching the Handicapped, Indiana University, February, 1973.

Martikean, A. The levels of questioning and their effects upon student performance above the knowledge level on Bloom's taxonomy of educational objectives (Field Research and Development Study - E585). Gary, IN: Indiana University Northwest, February, 1973.

Millett, G. B. Comparison of four teacher training procedures in achieving teacher and pupil "translation" behaviors in secondary school social studies (Doctoral dissertation, Stanford University, 1967).

Rogers, V. M. & Davis, O. L. Varying the cognitive levels of classroom questions: An analysis of student teachers' questions and pupil achievement in elementary social studies. Paper presented at the meeting of the American Educational Research Association, Lexington, 1970.

Ryan, F. L. The effects on social studies achievement of multiple student responding to different levels of questioning. Journal of Experimental Education, 1974, 42, 71-75.

Ryan, F. L. Differentiated effects of levels of questioning on student achievement. Journal of Experimental Education, 1973, 41, 63-67.

Savage, T. V. A study of the relationship of classroom questions and social studies achievement of fifth grade children (Doctoral dissertation, University of Washington, 1972).

Tyler, J. F. A study of the relationship of two methods of question presentation, sex, and school location to the social studies achievement of second grade children (Doctoral dissertation, University of Washington, 1971).

Walberg, H. J. Generalized regression models in educational research. American Educational Research Journal, 1971, 8, 71-91.